Learning about Activities and Objects from Video

Tony Cohn
What does an agent need to know about the world?

• What kind of objects there are.
• What they do/can be used for.
• What kinds of events there are.
• Which objects participate in which events.

... 

• How can an agent acquire this knowledge?
• How should it represent it?
Today’s talk

• Learning about
  - events: analyse activities in terms of event classes involving multiple objects
  - object categories via activity analysis

• Relational approach
This approach has a long history…

**Barrow and Popplestone:**
Relational descriptions in picture processing
*Machine Intelligence* 6, 1971

Relational descriptions of object classes + supervised learning

(re-)Connecting Logic and vision
(Kanade IJCAI’03)

From pixels to symbols to understanding
‘...let us consider the object recognition program in its proper perspective, as part of an integrated cognitive system. One of the simplest ways that such a system might interact with the environment is simply to shift its viewpoint, to walk round an object. In this way more information may be gathered and ambiguities resolved ..... 

...... Such activities involve planning, inductive generalization, and, indeed, most of the capacities required by an intelligent machine. To develop a truly integrated visual system thus becomes almost co-extensive with the goal of producing an integrated cognitive system.’

Barrow and Popplestone, 1971.
Object detection in the context of activity analysis

Movement can be at least as important as appearance in what we perceive:

Not just movement, but spatial relations between objects over time.

Heider & Simmel, 1944
Qualitative spatial/spatio-temporal representations

- Complementary to metric representations
- Human descriptions tend to be qualitative
- Naturally provides abstraction
  - Machine learning
- Provide foundation for domain ontologies with spatially extended objects
- Applications in geography, computer vision, robotics, NL, biology...
- Well developed calculi, languages
Qualitative temporal representations

Allen’s interval calculus

13 jointly exhaustive and pairwise disjoint relations
Qualitative spatial representations

Region Connection Calculus (RCC8)

(mereo)topology
Many other qualitative spatial calculi
- Other topological calculi
- Direction
- Size
- Distance
- Orientation
Using an HMM to ‘smooth’ relations

Sridhar et al.,
*COSIT 2011*
Learning relations

\[ r_t^m = (d_t^m, \dot{d}_t^m) \]

\[ q_t^m \rightarrow q_{t+1}^m \]

\[ r_t^m \rightarrow r_{t+1}^m \]

\[ q_t^m \]
Quantisation of Relational Features

---

2 discrete states

6 discrete states

8 discrete states

10 discrete states

12 discrete states

16 discrete states
Representing interactions relationally

\[ holds(X, Y, P, I_1) \land holds(X, Y, PO, I_2) \land holds(X, Y, DR, I_3) \]
\[ \land meets(I_1, I_2) \land meets(I_2, I_3) \land before(I_1, I_3) \]

1. Objects
2. Spatial Relationships (x 3)
3. Allen’s Temporal Relationships (x 13)
Representing an entire video relationally
Steps towards Cognitive Vision (at Leeds)

System which learned traffic behaviours (ICCV’98, IVC’00)
  • Qualitative spatio-temporal models

Learning of qualitative spatial relationships (ECAI’02)
  • Allows domain specific distinctions to be learned

Reasoning about commonsense knowledge of continuity to improve tracking (ECAI’04)

Learning symbolic descriptions of intentional behaviours
  • Use ILP to induce rules for simple games (AIJ 2004, ECAI’04, …)

Learning Qualitative Spatio-temporal event class descriptions
  • Supervised (ECAI’10a, ILP-11)
  • Unsupervised (ECAI’08, AAAI’10, ECAI’10, STAIRS’10, COSIT-11)

Functional Object Categories from a Relational Spatio-Temporal Representation (ECAI’08)

Workflow Activity Monitoring using the Dynamics of Pair-wise Qualitative Spatial Relations (MMM-12) …
Problem: Understanding Activities

- Point a camera at a scene with complex activities where objects are interacting.
- We start our analysis after obtaining object tracks.
Problem: Understanding Activities

- Activities consist of **events**.

- Events are **goal directed interactions** between a subset of objects.

- Events are **patterns** – instances of event classes
  - but may be **hidden by noise/coincidental interactions**.

- Can we **learn events** from complex activities in an **unsupervised** way despite the **presence of noise/coincidences**?
Events and Event Classes

What is an event?

• a set of spatio-temporal histories
  - some set of objects *interacting* at a particular time
  - each event is *unique*

What is an event class?

• at some level of abstraction, events will have *similar* descriptions
  - *qualitative* spatio-temporal change
  - *frequent* occurrences of similar events
Mining event classes

- What do we mean by *interacting*?
  - *How many objects* involved?
- What do we mean by *similar*?
- How *frequent*?
- *Complete* object histories, or *partial*?
  - How to split?
- Distinguishing *simple from complex* events
- Distinguishing *between contemporaneous events*
- How to find event classes efficiently?
  - How to search?
**Interactivity**: an attention mechanism for isolating subsets of objects over intervals

All objects interact evenly over interval

\( \tau_6 \) doesn’t interact much

Interactions temporally extended
Definitions

An *interaction graph* is a sub-graph of the activity graph that contains all spatial/temporal relations between its objects over some time interval.

A *cover* $\Lambda$ for an activity graph $A$ is a subset of the interaction graphs that jointly cover the (relevant) nodes of the activity graph. Furthermore, each interaction graph is labelled as *event* or *coincidence*.

A *model* $\Theta$ defines a probability function over interaction graphs.
Activity graph and cover
Mixture distribution over interaction graphs

A similarity measure between graphs

\[ p(g | c_i) = \sum_{k=1}^{N} q_k^i \mathcal{K}_d(g, h_k) \]

where \( \sum_{k=1}^{N} q_k^i = 1 \)
In summary:

- Each event class is specified non-parametrically by a set of event graphs.
- Try to explain data by finding a set of event classes such that the tracks can be divided into sets of tracklets each of which obeys the spatio-temporal constraints of some event class.
- A good explanation:
  - explains as much as possible
  - minimizes number of event classes
    (classes will tend to have many instances, and will be `large’)
  - has event classes all of whose graphs are similar
  - has events which have high degree of tracklet interaction and low object sharing with other events.
Activity graph and cover
Find MAP interpretation using MCMC

MCMC moves

- Birth/Death
- Split/Merge
  - Interactions
  - Region Histories
- Change
  - Spatial Relationships
Evaluation in aircraft domain

24 aircraft turnarounds – 37 hours

Single viewpoint

Semi-automated tracking of the plane, trolley, carriage, loader, bridge, plane-puller

Discard class labels

Obtain RCC3 relations in image plane
Results

Discovered two classes

<table>
<thead>
<tr>
<th>Semantic category</th>
<th>True positives</th>
<th>False positives</th>
<th>False negatives</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1 (un)loading</td>
<td>14</td>
<td>4</td>
<td>6</td>
<td>78</td>
<td>70</td>
</tr>
<tr>
<td>Class 2 bridge-on-off &amp; plane-puller-on</td>
<td>16</td>
<td>7</td>
<td>6</td>
<td>70</td>
<td>80</td>
</tr>
</tbody>
</table>
Learning object classes from behaviour (not appearance)

Most computer vision work on learning object classes recognises objects from their appearance

Can we categorise objects by what they do, not what they look like?
Inducing a functional object taxonomy (ECAI-2008)

Form Boolean matrix of the role played by objects in each event class (+ partially generalised classes)

<table>
<thead>
<tr>
<th>Objects</th>
<th>(E_1)</th>
<th>(E_2)</th>
<th>(E_m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(o_1)</td>
<td>0 1 0 0 0</td>
<td>0 0 1 0</td>
<td></td>
</tr>
<tr>
<td>(o_2)</td>
<td>1 0 0 0 1</td>
<td>0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>(\vdots)</td>
<td>0 0 0 0 0</td>
<td>0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>(o_n)</td>
<td>1 0 0 0 0</td>
<td>0 0 0 0</td>
<td>0 0 0 0</td>
</tr>
</tbody>
</table>

Compress the rows (pattern for each object) using PCA
Obtain object taxonomy by hierarchical-clustering of the compressed rows
Emergent object classes:

*Toy kitchen hierarchy*
Emergent object classes: 
Aircraft domain
Semi-supervised event learning

Look what’s *happening* over *there*

- “Deictic supervision”

• Just specify a rough spatio-temporal region for positive examples
  
  - No need to specify *exactly* which objects are involved in the event.

• We have developed a *transactional, typed* Inductive Logic Programming (ILP) system to induce rules.
What is Inductive logic programming?

- Machine learning, where the hypothesis space is the set of all logic programs
- Logic programs are a subset of First Order Logic
- A set of rules of the form:
  \[ \text{Event}(\ldots) \leftarrow \text{Condition}_1(\ldots) \land \ldots \land \text{Condition}_n(\ldots) \]
- Very expressive
- Learning consists of finding a set of rules such that all (most?) of the examples are correctly labelled by these rules.
- We use types to:
  - improve efficiency
  - reduce overgeneralisation from noisy examples
Evaluation

• 15 aircraft turnarounds
• 50,000 frames each turnaround
• 6 camera views
• Obtain tracks on 2D ground-plane
• ~350 spatial facts/video +temporal
• 10 event classes, 3-15 examples for each
• Many errors:
  - false/missing/displaced objects
  - broken/switched tracks
• Generate spatial relations between objects/IATA-zones
• Prolog rules determining temporal relations are in Background
• Leave-one-out (from turnarounds) testing
A Learned Event Model:

- \texttt{aircraft\_arrival([intv(T1,T2),intv(T3,T4)])} ←
  \texttt{surrounds(obj(aircraft(V)), right\_AFT\_Bulk\_TS\_Zone, intv(T1,T2))},
  \texttt{touches(obj(aircraft(V)), right\_AFT\_Bulk\_TS\_Zone, intv(T3,T4))},
  \texttt{meets(intv(T1,T2),intv(T3,T4))}.
Applying the learned rules:
## Results

<table>
<thead>
<tr>
<th>Event</th>
<th># examples</th>
<th>Learned rules</th>
<th>Hand-crafted rules</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>precision</td>
<td>recall</td>
</tr>
<tr>
<td></td>
<td></td>
<td>precision</td>
<td>recall</td>
</tr>
<tr>
<td>FWD_CN_LoadingUnloading_Operation</td>
<td>5</td>
<td>0.71</td>
<td>0.3</td>
</tr>
<tr>
<td>GPU_Positioning</td>
<td>4</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>Aircraft_Arrival</td>
<td>15</td>
<td>0.15</td>
<td>0.06</td>
</tr>
<tr>
<td>AFT_Bulk_LoadingUnloading_Operation</td>
<td>12</td>
<td>0.83</td>
<td>0.11</td>
</tr>
<tr>
<td>Left_Refuelling</td>
<td>6</td>
<td>0.38</td>
<td>0.5</td>
</tr>
<tr>
<td>PB_Positioning</td>
<td>15</td>
<td>0.25</td>
<td>0.5</td>
</tr>
<tr>
<td>Aircraft_Departure</td>
<td>10</td>
<td>0.33</td>
<td>0.14</td>
</tr>
<tr>
<td>AFT_CN_LoadingUnloading_Operation</td>
<td>7</td>
<td>0.54</td>
<td>0.4</td>
</tr>
<tr>
<td>PBB_Positioning</td>
<td>15</td>
<td>0.92</td>
<td>0.05</td>
</tr>
<tr>
<td>FWD_Bulk_LoadingUnloading_Operation</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Summary/novelty

- From pixels to symbolic, relational, qualitative behaviour/event descriptions
- Minimal supervision
- Multiple objects, shared objects, multiple simultaneous events,
- Robust computation of qualitative relations via HMM
- Functional object categorisation through event analysis

See papers for related work discussion

www.comp.leeds.ac.uk/qsr/publications.html
Research challenges/ongoing work

- New domains, longer scenes
  - Cognito project: learning workflows
  - Mind’s Eye project: spatio-temporal semantics of verbs
- Further experimentation with different sets of spatial relations
- Use induced functional categories to supervise appearance learning
- Learning probabilistic weights for rules (MLN)
- Interleaving induction/abduction to mitigate noise
- Cognitive evaluation of event classes and functional categories
- Learning a global model
  - temporal sequencing of event classes
- Online learning and Ontology alignment
- Language (+ vision)
- …
Any Questions?

Thanks to:

EPSRC, EU (CoFriend, Cognito), DARPA (Mindseye/Vigil)

David Hogg, Krishna Sridhar, Sandeep Dubba, QSR and CV groups at Leeds